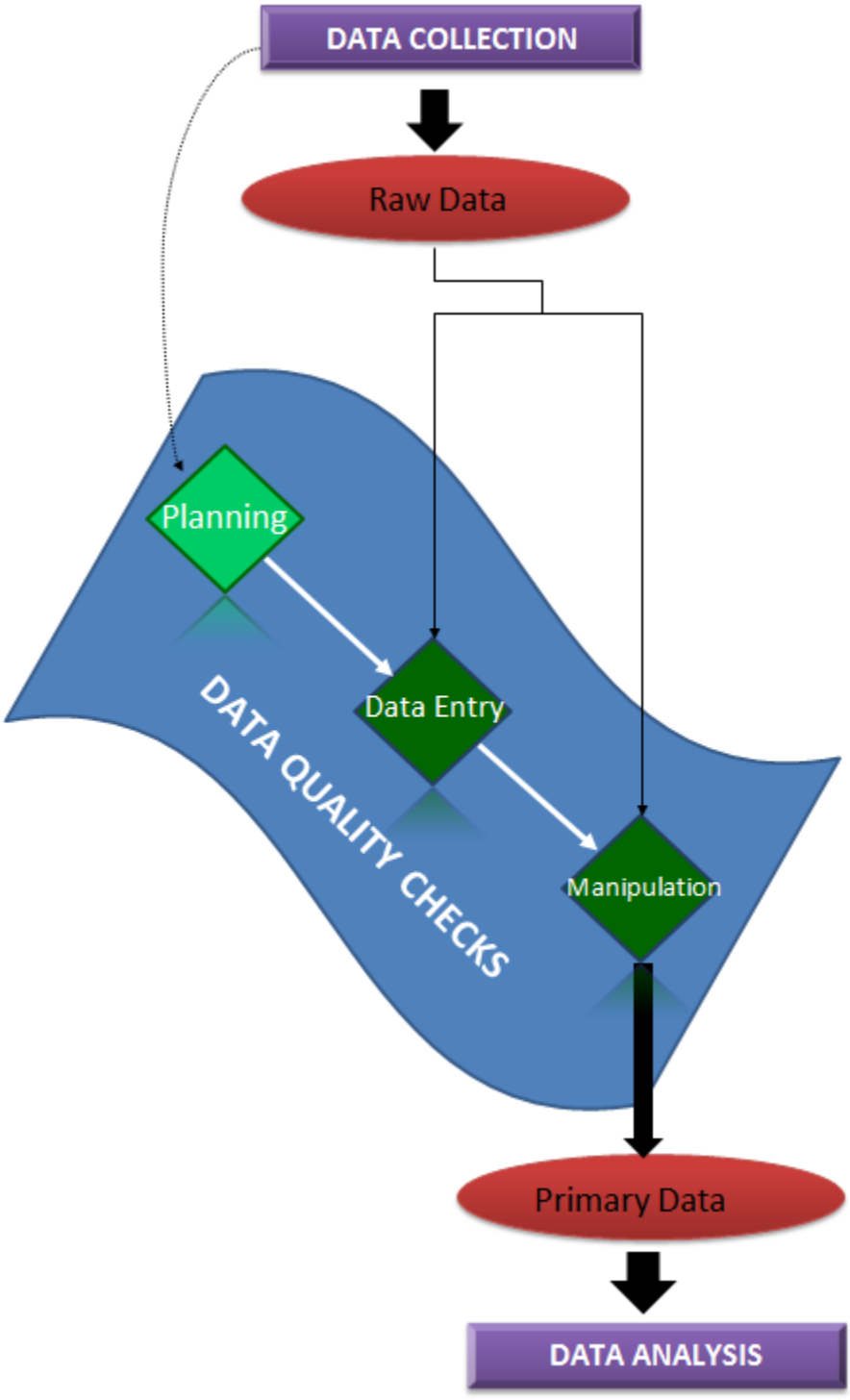
Transition from Raw to Primary Data

March 2018

Introduction

Raw data encompasses everything that was obtained in the data collection process, whether it is data directly gathered by the study or data that has been collected prior to the study that becomes part of the study datasets.

Data can come in a wide variety of formats; some immediately useful and ready for analysis whilst others will require steps to input, format, validate and derive before any analysis can be done. The result of this process is the primary data.



Overall Process

The components of the process of transferring from raw to primary data are:

- **Planning:** Specifying the format of how each individual piece of data will be stored (whether data are: numeric, coded, categorical, free text, identifier data, etc.), as well as which software will be used for entry and storage of the electronic data, and in the case of digital data collection, the data collection itself. Some of this may already be in place from the tool design, for example the responses to a question in a survey are likely to have been coded during the questionnaire design stage.
- **Data Collection/Entry:** Ensuring all data is available in electronic form. This is done during data collection where mobile devices are used but afterwards for paper questionnaires.
- **Manipulation:** Deriving additional information from the data and bringing all data sources together. For digital data collection this would include removing unneeded calculations and notes etc.
- **Data Quality Checks:** These checks should be taking place throughout the whole data transition process.

This document is primarily concerned with the elements of the Data Collection/Entry and Manipulation stages. The planning and data quality checking stages are outlined in more detail in other documents.

Digital Data Collection

Digital data collection is now becoming the norm and the process is much quicker than using paper questionnaires as the data collection and data entry stages of a project are effectively combined. Although there are potential problems due to technical issues, nearly all such problems can be dealt with by careful planning and training (e.g. regularly backing up data, carrying spare parts/batteries/additional devices, etc.) With digital data collection, the data entry system must be prepared and thoroughly checked in advance of the fieldwork. The design work is therefore moved to earlier in the project lifecycle; when using paper questionnaires, the data entry system can be designed during the fieldwork. Double data entry, which we would always suggest for paper questionnaires, can obviously not be done with digital data collection so there is more reliance on the enumerators to enter the data correctly during the fieldwork. Enumerators need to have the ability both to conduct the interviews and use the data entry system on the mobile devices. As mentioned earlier, these issues can be dealt with by careful planning and training.

Data Entry

Where raw data is not already in a digital format (e.g. paper questionnaires), it will require a data entry process to transfer the collected data into computerised format. The data entry process should ensure that all data stored digitally is a complete reflection of the raw data being entered.

We would suggest using software such as Microsoft Access or CSPro to create a data entry system with screens that resemble the questionnaire as much as possible. The system should be carefully designed so that it meets as many of the objectives for data storage and data quality checks as possible, whilst remaining intuitive for the data entry staff to use.

To help ensure the accuracy of the data we recommend using double data entry. The expected human error rate for a simple data input process, using only coded and numeric fields, has been shown in studies to be between 0.3% and 2% depending on the prior data entry experience of the individual. Although this may seem small, in the context of a moderately sized study with 50 pieces of information for each of 100 individuals, it corresponds to a minimum of 15 errors. Researchers have found that this can lead to a large effect on the eventual analytical findings potentially showing statistically significant results where the true values would not have done so and vice versa[1].

The probability of the same error being made independently by two people is very small (around 1 in 40,000 assuming an individual error rate of 0.5%). This means that the majority of data entry errors can be found by comparing the two sets of entered data and investigating instances where they differ by checking with the original copy to find the correct value. Errors often occur as a result of data not being recorded clearly or legibly or perhaps a confusing data entry system. Checks should be in place during the data collection process and in the design of the data entry system to prevent errors in the data.

Inputting data in coded form is much more preferable to typing out full responses. It will not only decrease the length of time taken per record, but it will also reduce the overall error rate. Obviously, many fields cannot be coded so must be entered in full, e.g. names, addresses and comments.

Manipulation

Digital Data Collection

As already mentioned digital data collection effectively combines the data collection and the data entry phases. However, the resulting data file will often need a great deal of manipulation before it can be used for analysis. With ODK for example, whenever you have an acknowledgement or a note in the form design, this will produce a separate column in the resulting data file, numerical codes will be stored as text, and multiple response questions will result in an initial column which lists the TRUE responses and a column for each possible option containing TRUE or FALSE as appropriate. The column headers will not only contain the variable name, but also the groupings contained in the system. Data from repeat groups will be stored on separate sheets in the resulting Excel file and a column headed "_parent_index" will identify the parent record in the main sheet.

Detailing how to deal with such data is beyond the scope of this document, but it is a good idea to create manipulation routines (e.g. VBA macros in Excel, R code) to convert the data into a usable format. These routines can then be used regularly for monitoring purposes throughout the fieldwork as new data is uploaded to the aggregate server.

[1] Barchard, K. A., & Pace, L. A. (2011, September). Preventing human error: The impact of data entry methods on data accuracy and statistical results. *Computers in Human Behaviour, 27*(5), 1834-1839.
Kawado, M. (2003, May). A Comparison of error detection rates between the reading. *Controlled Clinical Trials, 24,* 560-569

Paper Questionnaires

When using paper questionnaires there will also be some manipulation to be done after the data entry. The software used for entering the data is generally not the same software that will be used for analysis. Whatever software is used for data entry, data at different levels (e.g. data on individuals, household level data) should be stored in separate data files. Data at the lower level (e.g. the individual level) must contain relevant ID variables so that it can be linked to the data at the higher level.

Deriving Variables

Additional numerical fields can be derived as part of the manipulation process and stored alongside the original data. For example, converting areas recorded in different units into hectares, calculating age from date of birth, or calculating a poverty index from a combination of other variables. Where these variables can be derived it is better to rely on computed calculations for these values rather than on values calculated by hand and included as part of the raw data.

Categorical variables can also be derived from the data, for example, age could be split into groups or the responses from several fields could be combined to create a pass/fail type response.

Full documentation on how additional variables were derived should be included as part of the metadata. The inclusion of syntax files containing the commands for calculating the derived variables, in the data archive, is highly recommended.

Deriving Datasets

In addition to deriving individual variables, it will sometimes be necessary to derive complete datasets. For example, you might need to change the level at which the data are stored for a particular analysis. For example, climate data is generally presented on a day to day level, but it is often of more use to studies to derive monthly summaries to be used in analysis. Another example might be where the analysis is focused on the change from baseline to end-line and a derived dataset could be created containing the original baseline data alongside the new values. In both of these cases a whole dataset will be derived from the raw data and/or historic/external data. The data should be stored following the data storage recommendations and any calculations underlying the derivation should be documented in the metadata. In particular it should be clear how missing values were treated in the derivation process.

Data Quality Checks

Data checks should take place throughout the data collection and the transition of raw to primary data. Checking the data at the collection and entry stages helps to minimise errors before manipulations are done. Checking again after manipulation confirms that the manipulation was done correctly.

Summary

To summarise, there are a number of processes that need to be followed before data is ready for analysis. Remember also that it is the primary data and not generally the raw data that is put into the public domain, though the raw data should be kept in the internal project archive.

Associate Videos

Videos accompanying the original release of the CCAFS Data Management Support Pack in 2013 are available as a playlist on the Statistical Services Centre YouTube Channel at https://www.youtube.com/channel/UCs7EU95YMjhvNozJKCD92xQ/playlists. These videos have not been updated since the original release but are mostly still relevant.

In particular the playlist includes a video on Transition from Raw to Primary Data which is available at:
https://www.youtube.com/watch?v=IR0hbPIn_Yk&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpj&index=17